Journal of Engineering and Computer Innovations Vol. 2(2), pp. 22-27, February 2011
Available online at http://www.academicjournals.org/JECI
ISSN 2141-6508 ©2011 Academic Journals

Full Length Research Paper

Clustering analysis: A case study of the environmental data of RAMA-Toluca

Miguel Sánchez Sotelo[1], Rosa María Valdovinos Rosas[1]*, Roberto Alejo Eleuterio[2], Edgar Herrera[3] and Eduardo Gasca[4]

[1]Universidad Autonóma del Estado de México, Centro Universitario Valle de Chalco, Hermenegildo Galena No.3, 56615, Valle de Chalco, México.
[2]Tecnológico de Estudios Superiores de Jocotitlan, Carr. Toluca-Atlacomulco Km 44.8, 50700, Jocotitlan, México.
[3]Instituto Nacional de Investigación Nuclear ININ, Carr. Mexico-Toluca s/n, 52750, La Marquesa, Mexico.
[4]Instituto Tecnológico de Toluca, Av. Tecnológico s/n, 52140, Metepec, Mexico.

Accepted 7 January, 2011

Recently, the climatic analysis has been widely studied with artificial intelligence tools. The importance of this topic is based on the environment impact produced for natural variations of the data on a certain ecosystem. In this paper, a first study of the meteorological parameters obtained with the Automatic Network of Atmospheric Monitoring (by its abbreviation in Spanish, RAMA) of Toluca, Mexico, is exposed. The study period is from 2001 to 2008. RAMA-Toluca includes seven monitoring stations located in the Toluca Valley. Using clustering algorithms, the experimental results establish the base for determining the days of distribution in clusters, which could be oriented to the natural cluster that the days have in climatic seasons. However, the results show a different situation than the awaited one. With this, the bases for future work are in the climatic analysis context in Toluca Valley.

Key words: Clustering analysis, environmental data, RAMA-Toluca, days of distribution.

INTRODUCTION

The environmental data analysis is one topic that, in the last decades, has had importance in the scientific community. In this scope, studying the climatic change is the main environmental problem. The impact of this change has been foreseeable on the hydric resources, the productive ecosystems, the bio-diversity, the infrastructure, the public health and generally, on the diverse components included in the development process (Staines, 2007), which threatens the healthy environment and the quality of life.

In the Mexican State, particularly, in the metropolitan zone of the Toluca Valley (MZTV), it is possible to see that, when a rural place has been over-passed to an industrialized place, due to the continuous process of urbanization, the natural resources are devastated, and several environmental problems, like: bad use of the ground and reduction of the agricultural and forest border, invasion of protected natural areas, deforestation, erosion processes, forest fires, residues burnt in open-cast, pollution emissions by industries and damaged vehicles, are found (http://www.edomex.gob.mx/medioambiente).

For this reason, several artificial intelligence (AI) techniques are proposed to discover and conduct patterns of climate parameters in the MZTV. The AI is a discipline for developing software and hardware which can emulate the human actions, for example, manipulation of knowledge, generating conclusions, explaining the human reasoning and conducting it as if it was a human.

Clustering is the generic name of a great variety of techniques, useful for finding non-obvious knowledge in large data sets (Kotsiantis and Pintelas, 2004). There are two technique groups: The non-hierarchic techniques or the partition one and the hierarchic techniques. The first

*Corresponding author. E-mail: li_rmvr@hotmail.com.

Abbreviations: RAMA, by its abbreviation in Spanish, Automatic Network of Atmospheric Monitoring; **AA,** Adaptive Algorithm; **TC,** Toluca Center; **SL**, San Lorenzo Tepatitlan; **SM,** San Mateo Atenco.

one separates the data set in k groups, and the second one forms a set of several differentiation levels (MacKay, 2003). We can find different useful methods for determining the quality of clusters (Bolshakova et al., 2005). These methods use numerical measures on the clustering results by inferring the quality and describing the situation of a certain pattern inside the cluster.

Several studies are developed for handling this problem. Some studies, in which the climatic change was studied, were: Secretaria del medio ambiente (2007) and Parra-Olea et al. (2005). In general, the proposals consider a regional study of the climatic changes (Travasso et al., 2008) for projecting, regionally, the global predictions of the climate models available and to identify the effects of these changes (Gutierrez and Pons, 2006; Tebaldi and Knutti, 2009). On the other hand, there are several researchers that use either data mining (Steinbach et al., 2002; Atem et al., 2004) or clustering methods in different ways. For example, for discovering ecosystem patterns (Steinbach et al., 2001; Kumar et al., 2001), and improving the algorithm behavior (Gutiérrezr and Rodríguez, 2004), proposals were made on the weighted clustering method for analyzing infrequent patterns, or extreme events in the weather forecasts.

Based on the exposed patterns given previously, the object of this study was to analyze and discover the information that was inside the data bases provided by RAMA-Toluca. In particular, we analyzed the meteorological variables using clustering algorithms, for identifying the grouping in each year of the studied period (2001 – 2008). That is to say, we can know the distribution of the days between the groups and, in consequence, the seasons identified by the clustering methods. The paper is organized as follows: The clustering methods used in the study are exposed, followed by a description of the cluster validation algorithms which allow a corroboration of the group quality. Then the study zone and the meteorological parameters evaluated are given in detail, after which the experimental results are shown. Finally, the concluding remarks and the open lines of study are given.

CLUSTERING METHODS

The clustering process consists of a division of the data set in groups with similar objects. For measuring the similarity between objects, usually we use different distance measures, which are subsequently described in this work.

Adaptive algorithm

The adaptive algorithm (AA) is an incremental heuristic method which uses two parameters: distance threshold for creating groups (*t*) and a T fraction which determines the total confidence (⍰). The main function of the algorithm is to create groups based on t (weighted by ⍰). However, the first group settles down arbitrarily. The main processes of the AA are the following (Bow, 1992):

(i) The first group is determined arbitrarily.

(ii) When a sample was assigned to a certain group, the cluster center must be recalculated. This process can show that some samples change the cluster.

(iii) It is possible that the samples of a certain cluster change due to the iterative process.

(iv) The algorithm ends when there are no reassignments. At this time, the partition is considered stable.

K-means algorithm

K-means is a partition algorithm. In this way, similar samples are in the same cluster, and dissimilar samples are in different clusters (MacKay, 2003). In the process, the algorithm needs to define a unique parameter k. K defines the number of groups that will be found in the data set. For this, the K-means uses an iterative process, which starts by defining a sample prototype (centroid) as a cluster representative and is defined as the average of their samples. Next, the sample is assigned to the close centroid using a metric, commonly known as the Euclidean distance. Later, the centroid is recalculated using the new group formed. This process continues until a criterion is obtained, for example, the epochs number, no more replacements, etc (Garre et al., 2007).

The algorithm is faster and efficient; nevertheless, it has several limitations, such as, the a-priori knowledge about the cluster number inside the data set.

Validation algorithms

The cluster analysis consists of the clustering result evaluation, in order to find the partition that better fits the data (Halkidi et al., 2001). When the conglomerates were created, we needed to verify their quality through validation of algorithms (Bolshakova et al., 2005).

Cohesion

The cohesion can be defined as the sum of the proximities regarding the prototype (centroid) of a cluster (Bolshakova et al., 2005). The cohesion is given by:

$$Cohesion\ (Ci) = \sum_{x \notin C_i} proximity\ (x, C_i) \qquad (1)$$

Where *x* is the sample contained in cluster *i*; *Ci*, is the centroid of cluster *i*; and proximity is the squared Euclidean distance.

Separation

The separation between two clusters can be measured by the proximity of the prototypes (centroids) of two clusters. The separation is given by the next equation:

$$Separation\ (Ci) = proximity\ (C_i, C) \qquad (2)$$

Where Ci is the centroid of cluster *i*; *C*, is the general prototype (centroid) and *proximity* can be any metric (Tan et al., 2006).

Silhouette coefficient

This method combines two methods, which are cohesion and

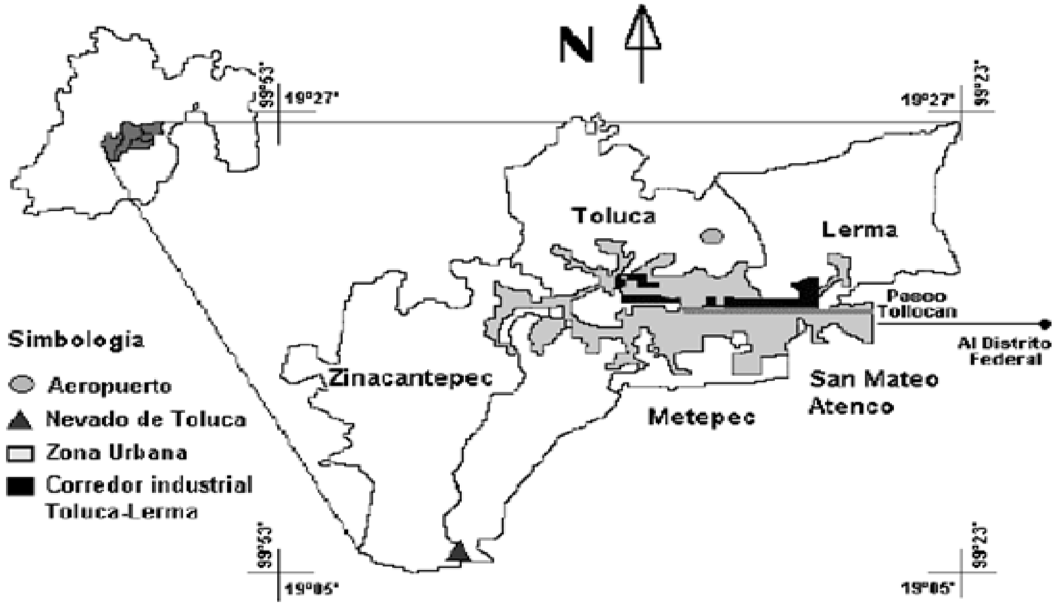


Figure 1. ZMVT Location (Extracted from Secretaria de medio ambiente, 2007).

Table 1. Samples by station and year.

Station	2001	2002	2003	2004	2005	2006	2007	2008	Total
CE	8627	8746	8528	8567	8104	6910	5546	2661	57689
SL	8171	8639	8692	8549	8441	7973	4087	7761	62313
SM	8213	8480	8478	8535	7909	8316	7828	595	58354

separation. The following steps explain the coefficient operation for a single object (Halkidi et al., 2001).

i.) For the *i*-'th object, the distance average is calculated for all objects that are in the same cluster, which is called value *ai*.
ii.) For *i*-'th object and any cluster that is empty, the distance average is calculated for all the objects in the next cluster. Finding the minimum value, regarding all clusters, is called *bi*.
iii.) For i-'th object, the silhouette coefficient is si = (bi - ai)/max(ai,bi). Where max(ai, bi) will be the maximum value between ai and bi.

The silhouette coefficient can vary between -1 and 1, and the maximum value of the coefficient is 1 when *ai* = 0. A negative value is undesirable because this corresponds to the case when *ai* is the average distance between the points in the cluster, and it is also greater than *bi*, which is the minimum average distance to the points of the other clusters. The best result desired is when the silhouette coefficient is positive (*ai* < *bi*) and when *ai* is close to 0.

To calculate the silhouette coefficient average (of one cluster), we take the coefficient average of all the points inside the cluster. A general measurement of a conglomerate can be obtained by calculating the silhouette coefficient average of all the points (Tan et al., 2006).

Study zone

In the ZMVT, the air quality has been measured since 1993 with 7 monitoring stations, and it includes seven municipalities in three zones which are shown in Figure 1. The monitoring stations store environmental data. For this research, the meteorological variables studied are: TMP (temperature), HR (relative humidity), PA (atmospheric pressure), RS (solar radiation), VV (wind speed) and DV (wind direction).

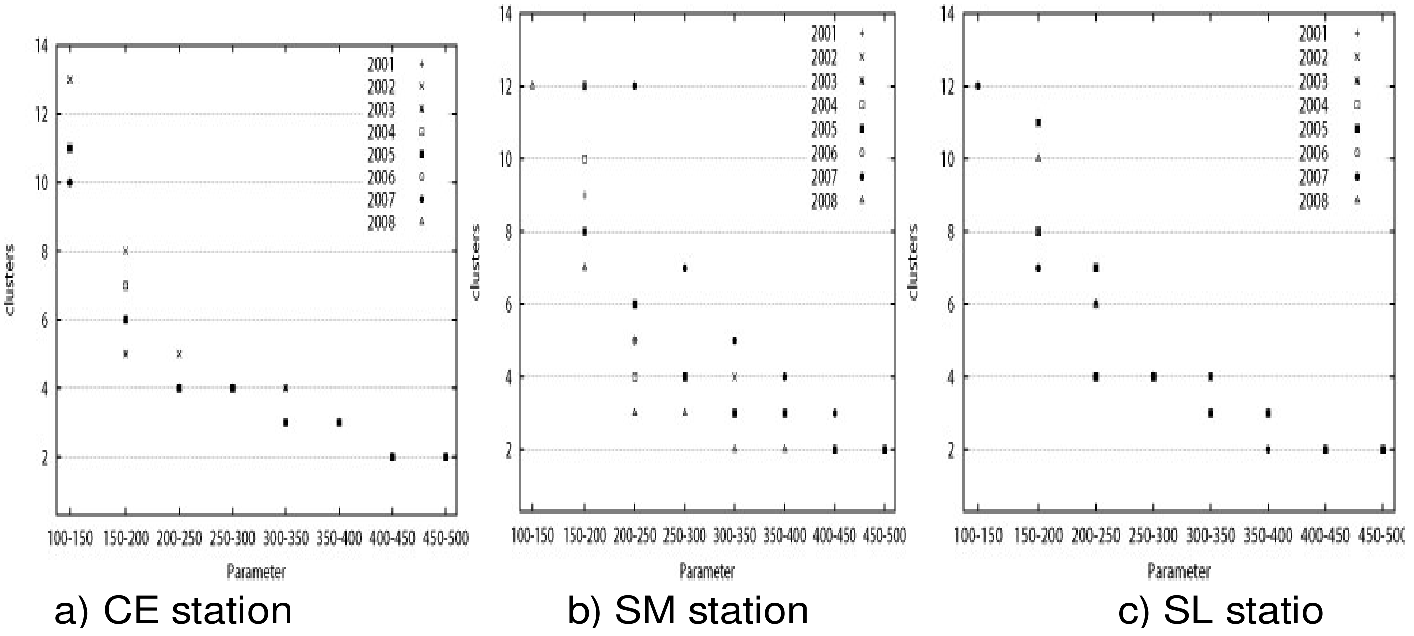
The data present several problems related to the monitoring station. These problems complicate their study, some or which are: faults in the sensor or its hard movements that are provoked by the wind or other causes. Some meteorological values are inconsistent with the reality (for example, a temperature of 80°C in winter) that the values are not captured completely (lost data). When the RAMA administrator identifies some of these problems, it marks the record for his later consideration.

In this way, the solution found was to choose the average value between the last and next real data for each feature. With this, we obtain a value in the real rank. On the other hand, when a sample loses more than 50% of the information, it is considered as noise and as such, it is eliminated.

The data, used for the study, were provided by 3 monitoring stations that showed different characteristics like: a great record number and a little lost of information. The monitoring stations are: Toluca Center (TC), San Lorenzo Tepatitlan (SL) and San Mateo Atenco (SM). Table 1 shows the number of samples by station and the number patterns per year in each station.

EXPERIMENTAL RESULTS

Here, the results of the clustering algorithms studied on



a) CE station

b) SM station

c) SL statio

Figure 2. Cluster number using the adaptive algorithm.

the data base provided by RAMA-Toluca are shown. Firstly, the data were filtered and, next, the clustering algorithms were applied: *k*-means and adaptive algorithm.

Some specifications for the k-means algorithm are the next: the initial seed was chosen randomly. The *k* value where k =2, 3, 4 for each data base (Table 1). On the other hand, for the adaptive algorithm, different thresholds by the data base were applied, and they were: 100 - 150, 150 - 200, 200 - 250, 250 - 300, 300 - 350, 350 - 400, 400 - 450, and 450 - 500, for the threshold and T-value, respectively. Figure 2 showed how the samples were grouped and how many groups were formed. The validity of the conglomerates quality, using the silhouette coefficient, is displayed in Figure 3. In Figures 2 and 3, it is possible to observe that when the clusters number is diminished, the quality is greater. This indicates that the best clustering is when the algorithm finds two clusters. On the other hand, Table 2 includes the samples grouped in each clustering results (in the case of two clusters).

The figures reflect the convergence existing between these two clustering algorithms, when both of them are found in the two groups. Regarding the groups' samples, it is possible to observe that one of the groups is bigger than the other one with almost a double quantity of the samples.

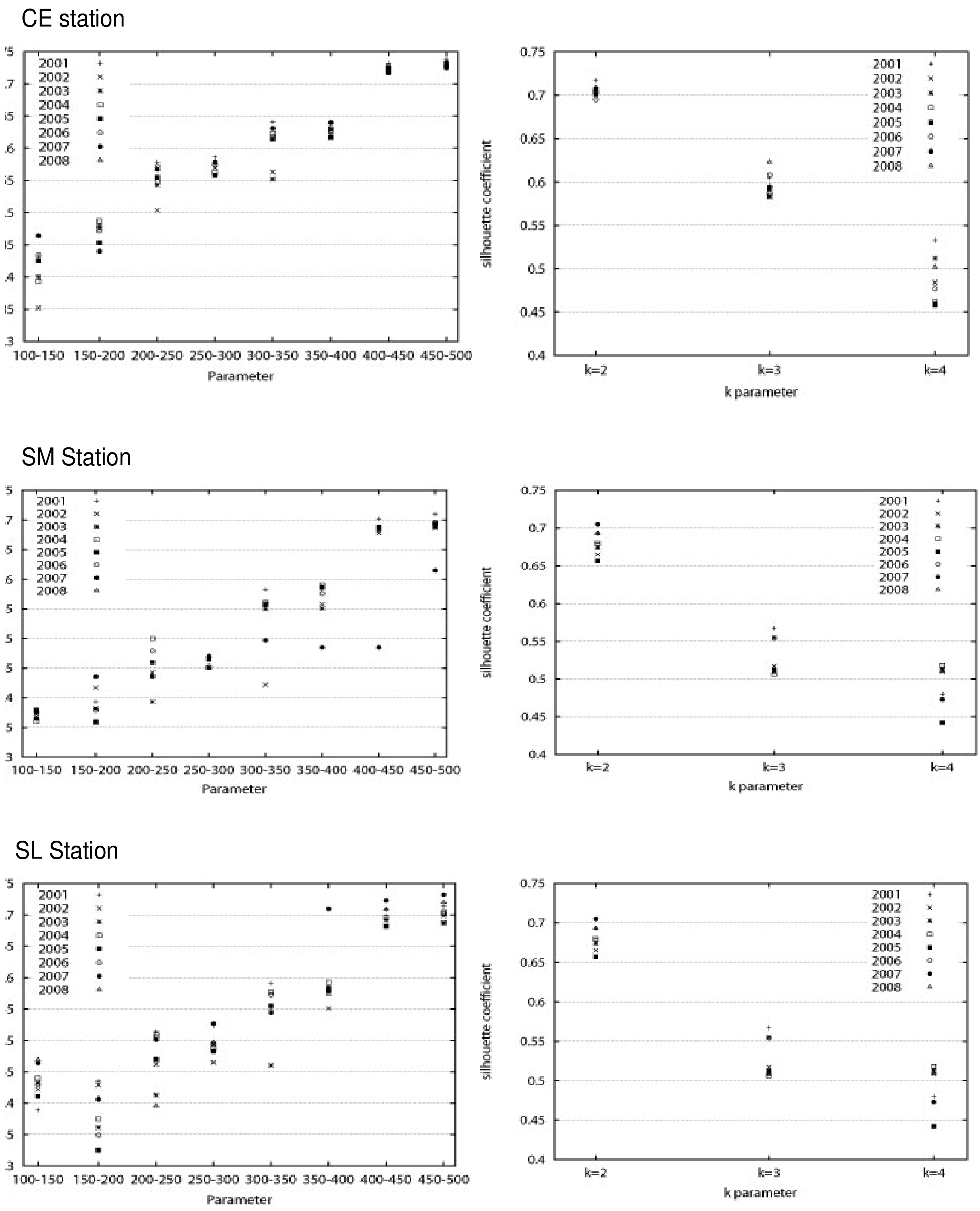
Conclusions

Throughout the year, the climate changes according to the cultural season. The hypothesis establishes that there are four seasons in one year. For this reason, we expect to find four conglomerates in the data set provided by RAMA-Toluca, because of the similarities between the samples of each season. Nevertheless, with the analysis exposed here, it was possible to identify that with the meteorological data analyzed, the clustering algorithms found only two great groups. In order to validate the quality of the clusters, the silhouette coefficient, cohesion and separation were used.

The preliminary results, exposed here, could indicate that in the meteorological data studied, the values of the samples of each year have similar features, mainly, of two seasons. In addition, due to the insignificant differences from each conglomerate, it is possible to suppose that, any climatic variation could happen before the year 2001.

With these results, it is possible to establish the bases of future works in this important topic, but several questions needs an answer: Is the cluster number equal to the season number? Is the behavior due to the climatic change? For answering these questions and obtaining a wider analysis, we came in contact with meteorological experts. During this time, we were in touch with the environmental engineering group of the Technological Institute of Toluca and Environment Secretariat in Toluca for improving the analysis. As such, we are sure that in a future work, we are going to expose the new analysis.

About the research in process, we are working with the unsupervised neural network SOM (Tan et al., 2006) for comparing several scenarios, for example, between 2001 and 2002 to 2008. The linear regression and correlation analysis is done by another study in process that would



CE station

SM Station

SL Station

Figure 3. The Silhouette coefficient. Figures from the left correspond to the silhouette coefficient of the adaptive algorithm, while figures from the right are the silhouette coefficient of the k-means algorithm.

Table 2. Samples grouped by cluster.

Algorithm	Year	SL		SM		CE	
		Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2
Adaptive	2001	6451	1719	6638	1575	6832	1795
	2002	6894	1745	6930	1550	6832	1795
	2003	6924	1768	6898	1580	6781	1747
	2004	6865	1684	6959	1576	6856	1711
	2005	6757	1684	6402	1507	6474	1630
	2006	6475	1498	6790	1526	5539	1371
	2007	3263	824	6215	1613	4419	1127
	2008	6305	1356	512	83	2067	594
K-means	2001	2207	5964	2241	5972	2193	6434
	2002	2294	6345	6294	2186	6572	2174
	2003	2346	6346	6233	2245	6387	2141
	2004	2306	6243	2272	6263	2141	6426
	2005	2254	6187	2116	5793	6057	2047
	2006	5953	2020	6161	2155	1845	5066
	2007	3098	989	6293	1535	1407	4139
	2008	5933	1728	427	168	753	1908

soon be finished.

The open lines point to the study of other data bases with information of more years and of the other states or countries. Also, it is possible to include the analysis of other years and other climatic data bases, as well as to use other algorithms such as ISODATA and DBSCAN (Martín et al., 1996). In the same way, we analyze the convenience of including other validation methods and studying methods for handling the lost data.

ACKNOWLEDGMENTS

This work has been partially supported by grants: INV-2010-XX (from the Spanish 2010-UJI promotion plan), Reactor operation UR-001 (from the ININ), PROMEP 103.5/09/4195 (from the Mexican SEP), 2933/2010U (from the UAEM) and SDMAIA-010 (from the TESJO project).

REFERENCES

Atem OI, Luengo F, Cofiño AS, Gutíerrez JM (2004). Grid oriented implementation of selforganizing maps for data mining in meteorology. In: In meteorology, in grid computing. Proc. of 1st European Across GRIDs Conference, pp. 163–170.

Bolshakova N, Azuaje F, Cunningham P (2005). An integrated tool for microarray data clustering and cluster validity assessment. Bioinformatics, 21(4): 451–455.

Bow ST (1992). Pattern Recognition and Image Preprocessing Marcel Dekker Inc.

Garre M, Cuadrado JJ, Sicilia MA, Rodriguez D, Rejas R (2007). Comparison of different algorithms from clustering in the cost estimation in the development of software. Revista Española de Innovación, Calidad e Ingeniería del Software, 3(1): 6–22.

Gutiérrezr JM, Cofiño AS, Cano R, Rodríguez MA (2004). Clustering methods for statistical downscaling in short-range weather forecasts. Monthly Weather Rev., 132(9): 2169-2183.

Gutierrez JM, Pons MR (2006). Numerical modeling of climate change: Scientific basis, uncertainties and projections for the iberian peninsula. Cuaternario Geomorfología, 20(2-4): 15-28.

Halkidi M, Batistakis Y, Vazirgiannis M (2001). On clustering validation techniques. J. Intelligent Inf. Syst., 17(2-3): 107–145.

Kotsiantis, SB, Pintelas PE (2004). Recent advances in clustering: A brief survey. WSEAS Transactions on Infor. Sci. Appl., 1: 73–81.

Kumar V, Steinbach M, Tan P, Klooster S, Potter C, Torregrosa A (2001). Mining scientific data: Discovery of patterns in the global climate system. In Proc. of the Joint Statistical Meetings.

MacKay DJC (2003). Information Theory, Inference, and Learning Algorithms. Cambridge University Press.

Martín E, Hans PK, Jörg S, Xiaowei X (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proc. of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), pp. 226–231.

Parra-Olea G, Martínez-Meyer E, Pérez-Ponce G (2005). Forecasting climate change effects on salamander distribution in the highlands of central Mexico. Biotropica, 37(2): 202–208.

Secretaria del medio ambiente (2007). Aire Limpio: Programa para el Valle de Toluca 2007-2011. Edomexico.

Staines F (2007). Cambio climático: interpretando el pasado para entender el presente. *Ciencia* Ergo Sum, 14(4): 345–351.

Steinbach M, Tan P, Vipin K, Klooster S, Potter C, Torregrosa A (2002). Discovery of climate indices using clustering. In: In Proc. of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 24–27.